

**Problem 1**

No, there is no such guarantee. There may be datapoints, which either (1) violate the margin, i.e.  $0 \leq \xi_i \leq 1$  or (2) are on the wrong side of the hyperplane, i.e.  $y_i(w^\top x_i + b) \geq 1 - \xi_i$  does not hold, in this case  $\xi_i > 1$ .

**Problem 2**

$C < 0$  we have from the Lagrangian of slack variables that  $0 \leq \alpha_i \leq C$  but this obviously does not hold. If  $C = 0$ , then we have no penalty term in the optimization problem and from the Lagrangian of slack variables it follows that

$$\begin{aligned} \alpha_i &= -\mu_i \\ 0 \leq \alpha_i &= -\mu_i \\ 0 \leq -\mu_i & \quad (\text{violation of dual feasibility of Lagrangian multipliers } \mu_i \geq 0) \end{aligned}$$

Thus  $C > 0$ .

**Problem 3**

We prove that  $K(x, y) = (x^\top y + c)^d$  is a kernel by the construction rules.

- $K_1(x, y) = x^\top y$  is a kernel by Rule 5, where  $B = I$ .
- $K_2(x, y) = c$  is a kernel by Rule 4, where  $\phi(z) = \sqrt{c}$ ,  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$ ,  $m = 1$  and  $K_3(x, y) = x^\top y$
- $K_3(x, y) = K_1(x, y) + K_2(x, y)$  is a kernel by Rule 1
- $K_4(x, y) = K_3(x, y)^d$  is a kernel by recursive application of Rule 3

Thus  $K(x, y) = (x^\top y + c)^d$  is a valid kernel.

**Problem 4**

No, we can not apply it directly to our data because the feature map is infinite dimensional. For this we would need to define a representation of the inner product of two infinite vectors  $\phi_\infty(x), \phi_\infty(y)$  as a function  $K(x, y) = \phi_\infty(x)^\top \phi_\infty(y)$ .

**Problem 5**

$$\begin{aligned}
K(x, y) &= \sum_{i=0}^{\infty} \phi_{\infty,i}(x) \phi_{\infty,i}(y) \\
&= \sum_{i=0}^{\infty} \frac{1}{\sqrt{i!}} e^{\frac{-x^2}{2\sigma^2}} \left(\frac{x}{\sigma}\right)^i \frac{1}{\sqrt{i!}} e^{\frac{-y^2}{2\sigma^2}} \left(\frac{y}{\sigma}\right)^i \\
&= \sum_{i=0}^{\infty} \frac{1}{i!} e^{\frac{-x^2-y^2}{2\sigma^2}} \left(\frac{xy}{\sigma^2}\right)^i \\
&= e^{\frac{-x^2-y^2}{2\sigma^2}} \sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{xy}{\sigma^2}\right)^i && \text{(by Taylor series of } e^z \text{)} \\
&= e^{\frac{-x^2-y^2}{2\sigma^2}} e^{\frac{xy}{\sigma^2}} \\
&= e^{\frac{-x^2-y^2+2xy}{2\sigma^2}} \\
&= e^{-\frac{(x^2+y^2-2xy)}{2\sigma^2}} \\
&= e^{-\frac{|x-y|^2}{2\sigma^2}} && \text{(by definition Gaussian Kernel)} \\
&= \phi(x)^\top \phi(y)
\end{aligned}$$

Yes, overfitting can occur if  $\sigma \ll 1$

**Problem 6****Problem 7**

$$\begin{aligned}
\|x - x^{(s_i)}\|_2 &= \sqrt{\sum_{j=1}^M (x_j - x_j^{(s_i)})^2} \\
&= \sum_{j=1}^M (x_j - x_j^{(s_i)})^2 && \text{(squared distance does not change k nearest neighbors of } x \text{)} \\
&= \sum_{j=1}^M (\phi(x_j) - \phi(x_j^{(s_i)}))^2 && \text{(by feature map } \phi(x) \text{)} \\
&= (\phi(x) - \phi(x^{(s_i)}))^\top (\phi(x) - \phi(x^{(s_i)})) \\
&= \phi(x)^\top \phi(x) - 2\phi(x)^\top \phi(x^{(s_i)}) + \phi(x^{(s_i)})^\top \phi(x^{(s_i)}) \\
&= K(x, x) + K(x^{(s_i)}, x^{(s_i)}) - 2K(x, x^{(s_i)})
\end{aligned}$$


---